

AD-A091 033

PITTSBURGH UNIV PA INST FOR STATISTICS AND APPLICATIONS F/6 6/3
DIVERSITY AND DISSIMILARITY COEFFICIENTS: A UNIFIED APPROCH.(U)
JUL 80 C R RAO F49620-79-C-0161

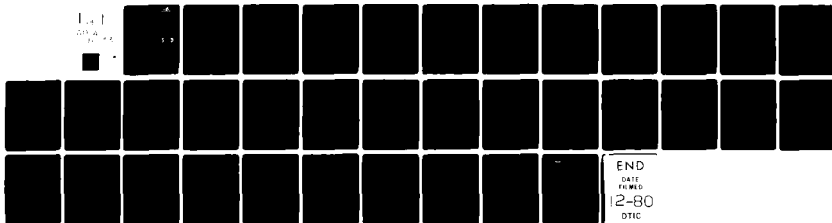
UNCLASSIFIED

TR-80-10

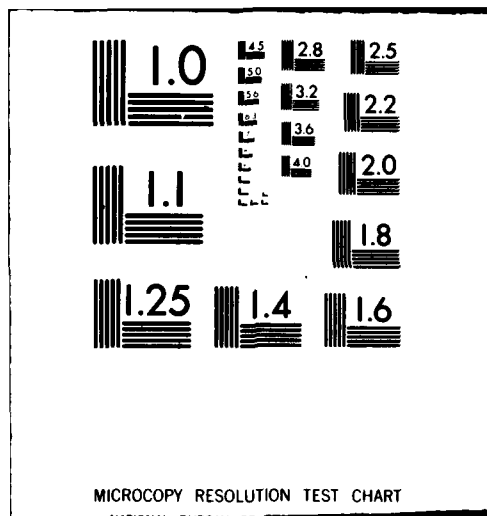
AFOSR-TR-80-0984

NL

1-1
00 00 00



END
DATE
FILMED
12-80
DTIC



AFOSR-TR- 80 - 0984

④ LEVEL II

AD A091033

DIVERSITY AND DISSIMILARITY COEFFICIENTS;

A UNIFIED APPROACH

C. Radhakrishna Rao*

DTIC
ELECTE
OCT 31 1980
S B D

July 1980

Technical Report No. 80-10

—Institute for Statistics and Applications
Department of Mathematics and Statistics
University of Pittsburgh
Pittsburgh, PA. 15260

DDC FILE COPY

*The work of this author is sponsored by the Air Force
Office of Scientific Research under Contract F49620-79-0161.
Reproduction in whole or in part is permitted for any purpose
of the United States Government.

80 10 6 099

DISTRIBUTION STATEMENT A

Approved for public release;
Distribution Unlimited

DIVERSITY AND DISSIMILARITY COEFFICIENTS:

A UNIFIED APPROACH

C. Radhakrishna Rao*

Summary: Three general methods for obtaining measures of diversity within a population and dissimilarity between populations are discussed. One is based on an intrinsic notion of dissimilarity between individuals and others make use of the concepts of entropy and discrimination. A criterion is developed for choosing a measure of diversity in a given class of measures. The Gini-Simpson index of diversity is derived as the solution to a functional equation.

AMS Classification 62H30

Key Words: Entropy, Information, Diversity, Dissimilarity, Similarity, Mahalanobis D^2 , Size and Shape Factors, Discrimination, Geodesic Distance

***The work of this author is sponsored by the Air Force Office of Scientific Research under Contract F49620-79-0161. Reproduction in whole or in part is permitted for any purpose of the United States Government.**

**AIR FORCE OFFICE OF SCIENTIFIC RESEARCH (AFSC)
NOTICE OF TRANSMITTAL TO DDC
This technical report has been reviewed and is
approved for public release IAW AFR 190-12 (7b).
Distribution is unlimited.
A. D. BLOSE
Technical Information Officer**

1. INTRODUCTION

There is an extensive literature on measures of diversity within populations and dissimilarity or similarity between populations. They have been used in a wide variety of studies in anthropology (Rao, 1948; Mahalanobis, Majumdar and Rao, 1949; Majumdar and Rao, 1958; Rao, 1971a,b, 1977), in genetics (Cavalli-Sforza, 1969; Karlin et al, 1979; Morton and Lalouel, 1973; Nei, 1978; Sanghvi, 1953; Sanghvi and Balakrishnan, 1972), in economics (Gini, 1912; Sen, 1973) in sociology (Agresti and Agresti, 1978) and in biology (Sokal and Sneath, 1963; Pielou, 1975; Patil and Taille, 1979). A complete bibliography of papers on measures of diversity and their applications is compiled by Dennis et al (1979).

Most of these measures are based on heuristic considerations; some are derived from mathematically well postulated axioms, while others are constructed using possible models for genetic and environmental mechanisms causing differences between individuals and populations. The object of this paper is to review some of these measures and to provide some unified approaches for deriving them.

We consider a set of populations $\{\pi_i\}$ where the individuals of each population are characterized by a set of measurements $X \in (\Omega, \mathcal{B})$, a measurable space. The probability distribution function of X in π_i is denoted by P_i and the convex set generated by $\{P_i\}$ is denoted by \mathcal{P} . A diversity coefficient

(DIVC) is a mapping from P into the real line, which reflects differences between individuals (X 's) within a population. We denote the DIVC of π_i by H_i (the symbol H is used to indicate heterogeneity). A dissimilarity coefficient (DISC) or a similarity coefficient (SIMC) is a mapping from $P \times P$ into the real line, which reflects the differences or similarities between populations. We denote a DISC between π_i and π_j by D_{ij} and a SIMC by S_{ij} .

2. COEFFICIENTS BASED ON INTRINSIC DIFFERENCES BETWEEN INDIVIDUALS

2.1 General Theory

We start first by choosing a non-negative symmetric function $d(X_1, X_2)$ which is a measure of difference between two individuals with $X = X_1$ and $X = X_2$, without any reference to the probability distributions of X_1 and X_2 . The choice of $d(X_1, X_2)$ naturally depends on the nature of the practical problem under investigation. We define the DIVC (diversity coefficient) of π_i as

$$H_i = \int d(X_1, X_2) P_i(dX_1) P_i(dX_2) \quad (2.1.1)$$

i.e., as the average difference between two randomly drawn individuals from π_i . Suppose that one individual is drawn from π_i and another from π_j . Then the average difference is

| | |
|---|----------------------|
| By <input checked="checked" type="checkbox"/> | |
| Distribution/ <input type="checkbox"/> | |
| Availability Codes <input type="checkbox"/> | |
| Dist | Avail and/or Special |
| A | |

$$H_{ij} = \int d(X_1, X_2) P_i(dX_1) P_j(dX_2). \quad (2.1.2)$$

We expect H_{ij} to be larger than the average of H_i and H_j , in which case the DISC (dissimilarity coefficient) between π_i and π_j may be defined by what can be termed as the Jensen difference.

$$D_{ij} = H_{ij} - \frac{1}{2} (H_i + H_j). \quad (2.1.3)$$

The expression (2.1.3) will be non-negative for any i and j iff $d(X_1, X_2)$ is chosen such that the function H defined on \mathcal{P} as in (2.1.1) is concave. This can be easily verified by considering $P_0 \in \mathcal{P}$ where

$$P_0 = \lambda P_i + (1 - \lambda)P_j, \quad 0 < \lambda < 1 \quad (2.1.4)$$

and computing

$$\begin{aligned} H_0 &= \int d(X_1, X_2) P_0(dX_1) P_0(dX_2) \\ &= \lambda^2 H_i + (1 - \lambda)^2 H_j + 2\lambda(1 - \lambda)H_{ij}. \end{aligned} \quad (2.1.5)$$

Then

$$\begin{aligned} H_0 &- (\lambda H_i + (1 - \lambda)H_j) \\ &= 2\lambda(1 - \lambda)(H_{ij} - \frac{1}{2} H_i - \frac{1}{2} H_j) = 2\lambda(1 - \lambda)D_{ij}. \end{aligned} \quad (2.1.6)$$

The concavity of H ensures that $D_{ij} \geq 0$ and vice-versa.

2.2 Some Examples

(1) Let $X \in \mathbb{R}^m$, a real vector space of m dimensions

furnished with an inner product $(x,y) = x'Ay$, where A is a positive definite matrix. Define

$$d(X_1, X_2) = (X_1 - X_2, X_1 - X_2). \quad (2.2.1)$$

Let $X \sim (\mu_i, \Sigma_i)$ in π_i (i.e., X is distributed with mean vector μ_i and dispersion matrix Σ_i). Then

$$\begin{aligned} H_i &= 2 \operatorname{tr} A \Sigma_i \\ H_{ij} &= \operatorname{tr} A \Sigma_i + \operatorname{tr} A \Sigma_j + \delta'_{ij} A \delta_{ij} \end{aligned} \quad (2.2.2)$$

where tr stands for the trace of a matrix and $\delta_{ij} = \mu_i - \mu_j$. Applying the formula (2.1.3)

$$D_{ij} = \delta'_{ij} A \delta_{ij}. \quad (2.2.3)$$

If $\Sigma_i = \Sigma$ for all i and $A = \Sigma^{-1}$, (2.2.3) becomes the Mahalanobis D^2 between π_i and π_j .

(2) Let $X = (x_1, \dots, x_m)$ where x_i can take only a finite number of values. For instance x_i may stand for the type of gene allele at a given locus i on a chromosome. In such a case an appropriate measure of difference between two vectors X_1 and X_2 is

$$d(X_1, X_2) = m - \sum \delta_r \quad (2.2.4)$$

where $\delta_r = 1$ if the r -th components of X_1 and X_2 agree and zero otherwise. Let x_r take k_r different values with probabilities

$$p_{i1}, \dots, p_{ir}$$

in population π_i . Define

$$j_{ii}^{(r)} = E(\delta_r) = \sum_{s=1}^{k_r} p_{irs}^2 \quad (2.2.5)$$

when X_1, X_2 are independently drawn from π_i and

$$j_{ij}^{(r)} = E(\delta_r) = \sum_{s=1}^{k_r} p_{irs} p_{jrs} \quad (2.2.6)$$

when X_1 is drawn from π_i and X_2 from π_j . Then

$$\begin{aligned} H_i &= \sum_{r=1}^m (1 - j_{ii}^{(r)}) = m(1 - J_{ii}) \\ H_{ij} &= \sum_{r=1}^m (1 - j_{ij}^{(r)}) = m(1 - J_{ij}) \end{aligned} \quad (2.2.7)$$

$$D_{ij} = H_{ij} - \frac{1}{2} (H_i + H_j)$$

$$= m \left[\frac{1}{2} (J_{ii} + J_{jj}) - J_{ij} \right] = \frac{1}{2} \sum_{r=1}^m \sum_{s=1}^{k_r} (p_{irs} - p_{jrs})^2. \quad (2.2.8)$$

The expression (2.2.8) without the factor m has been called by Nei (1978) as "a minimum estimate of the net codon difference per locus" and used by him and his colleagues (see the list of references in Nei, 1978) as a measure of genetic distance in phylogenetic studies.

Note 1. When $m = 1$, we have a single multinomial and the expression (2.2.8) reduces to the Gini-Simpson index

$$1 - \sum_{i=1}^k p_i^2 \quad (2.2.9)$$

where p_1, \dots, p_k are the cell probabilities. [This measure was introduced by Gini (1912) and used by Simpson (1949) in biological work]. The properties of (2.2.9) have been studied by various authors (Bhargava and Doyle, 1974; Bhargava and Uppuluri, 1975; Agresti and Agresti, 1978).

Note 2. It is seen that H_i as defined in (2.2.7) depends only on the marginal distributions of x_i , $i = 1, \dots, m$, and is additive with respect to the characters examined. These properties arise from the way the difference function (2.2.4) is defined. The DISC (2.2.8) is specially useful in evolutionary studies as suggested by Nei (1978).

Note 3. We may consider the joint distribution of (x_1, \dots, x_m) as a combined multinomial with $k = k_1 \times \dots \times k_m$ classes and apply the formula (2.1.1) to measure diversity. In such a case the difference between two individuals takes the value 1 when all the components x_i agree and the value zero if at least one is different. This leads to an expression different from (2.2.8) as the basic function for assessing the differences between individuals is not the same. When x_1, \dots, x_m are independently distributed, an explicit expression for the DIVC based on the combined multinomial reduces to

$$H = 1 - [1 - H(1)] \dots [1 - H(m)] \quad (2.2.10)$$

where $H(r)$ is the DIVC based on x_r , the r -th character only. It may be noted that the expression for DIVC given in (2.2.7) is $H = \sum H(r)$ whether x_i are independently distributed or not.

2.3 Apportionment of DIV

With the DIVC as defined by (2.1.1) and using the concavity property, the DIV in a mixture of populations can be apportioned in a natural way as between and within populations. If P_1, \dots, P_k are the distributions of X in π_1, \dots, π_k and $\lambda_1, \dots, \lambda_k$ are the apriori probabilities, then the distribution in the mixture π_0 is $\lambda_1 P_1 + \dots + \lambda_k P_k$. It is easily seen that

$$\begin{aligned} H_0 &= \sum \lambda_i H_i + \sum \sum \lambda_i \lambda_j D_{ij} \\ &= H(w) + D(b) \end{aligned} \quad (2.3.1)$$

where $D_{ij} = H_{ij} - (H_i + H_j)/2$ is the DISC between π_i and π_j , $H(w)$, the DIV within populations, is the weighted average of the DIV's within populations and $D(b)$, the DIS between populations, is the weighted average of the DISC's between all pairs of populations. The ratio

$$G(b) = \frac{D(b)}{H_0} \quad (2.3.2)$$

is an index of diversity between populations.

Different choices of the difference function $d(X_1, X_2)$ may give different values to the ratio $G(b)$. In Section 3, we shall discuss this problem in a more general context.

Let us consider k populations as in example (1) of Section 2.2 where in π_i , the m -vector variable $X \sim (\mu_i, \Sigma)$ and choose $d(X_1, X_2)$ as the Mahalanobis D^2 (formula (2.2.3) with $A = \Sigma^{-1}$). Further let π_0 be a mixture of π_1, \dots, π_k with apriori probabilities $\lambda_1, \dots, \lambda_k$. Then using the expressions (2.2.2), the decomposition (2.3.1) becomes

$$\begin{aligned} H_0 &= H(w) + D(b) \\ &= 2m + \sum \sum \lambda_i \lambda_j \delta'_{ij} \Sigma^{-1} \delta_{ij} \\ &= 2m(1+V) \end{aligned} \quad (2.3.3)$$

where $\delta_{ij} = \mu_i - \mu_j$. Thus the diversity within populations is $2m$ and the ratio $G(b)$ of (2.3.2) is V which is the weighted combination of Mahalanobis D^2 's for all pairs of populations. The author has suggested (see Mahalanobis, Majumdar and Rao, 1949) the use of an estimate of V in the selection of variables to maximize dissimilarity between populations.

Let us consider example (2) of Section 2.2 and denote by π_0 , the mixture of π_1, \dots, π_k with apriori probabilities $\lambda_1, \dots, \lambda_k$. In this case (2.3.1) becomes, with J_{ij} as defined in (2.2.7),

$$H_0 = m[\sum \lambda_i (1 - J_{ii}) + \sum \sum \lambda_i \lambda_j (\frac{1}{2} J_{ii} + \frac{1}{2} J_{jj} - J_{ij})] \quad (2.3.4)$$

which is the decomposition obtained by Nei (1973) and Chakravorthy (1974). The ratio $G(b)$ defined in (2.3.2) is

$$G(b) = \frac{\sum \sum \lambda_i \lambda_j (\frac{1}{2} J_{ii} + \frac{1}{2} J_{jj} - J_{ij})}{1 - \sum \sum \lambda_i \lambda_j J_{ij}} . \quad (2.3.5)$$

The ratio (2.3.5) obtained by considering only the two populations π_i and π_j with equal prior probabilities

$$\theta_{ij} = \frac{J_{ii} + J_{jj} - 2 J_{ij}}{4 - J_{ii} - J_{jj} - 2 J_{ij}} \quad (2.3.6)$$

is the hybridity coefficient of Morton (1973) who used it as a DISC between π_i and π_j in phylogenetic studies.

2.4 Decomposition of DIVC and DISC

In the method outlined in Section 2.1, the basic expression which determines the DIVC and DISC is the difference function $d(X_1, X_2)$. Any decomposition of $d(X_1, X_2)$ such as

$$d(X_1, X_2) = d_1(X_1, X_2) + \dots + d_c(X_1, X_2) \quad (2.4.1)$$

provides us with a corresponding decomposition of the DIVC for π_i

$$H_i = H_i^{(1)} + \dots + H_i^{(c)} \quad (2.4.2)$$

where $H_i^{(s)} = E[d_s(X_1, X_2) | P_i]$, and of the DISC between π_i and π_j

$$D_{ij} = D_{ij}^{(1)} + \dots + D_{ij}^{(c)} \quad (2.4.3)$$

where $D_{ij}^{(s)}$ is obtained from $H_i^{(s)}$, $H_j^{(s)}$ and $H_{ij}^{(s)}$ using the formula (2.1.3).

Let $X \sim (\mu_i, \Sigma)$ in π_i and denote the eigen values of Σ by $\theta_1 \geq \dots \geq \theta_m$ and the corresponding eigen vectors by L_1, \dots, L_m . If we choose

$$d(X_1, X_2) = (X_1 - X_2)'(X_1 - X_2)$$

i.e., the simple Euclidean distance in R^m , then

$$d(X_1, X_2) = [L_1'(X_1 - X_2)]^2 + \dots + [L_m'(X_1 - X_2)]^2 \quad (2.4.4)$$

gives the decomposition of DIVC for π_i

$$H_i = 2 \operatorname{tr} \Sigma = 2 \theta_1 + \dots + 2 \theta_m \quad (2.4.5)$$

which is the familiar decomposition of total variability with respect to m characters in terms of principal components (Rao, 1964). The corresponding decomposition of DISC between π_i and π_j is

$$D_{ij} = \delta_{ij}' \delta_{ij} = (L_1' \delta_{ij})^2 + \dots + (L_m' \delta_{ij})^2 \quad (2.4.6)$$

where $\delta_{ij} = \mu_i - \mu_j$, the difference in the mean vectors for π_i and π_j . However, if we choose

$$d(X_1, X_2) = (X_1 - X_2)' \Sigma^{-1} (X_1 - X_2)$$

i.e., the Mahalanobis distance between two individuals then we have a different decomposition

$$D_{ij} = \delta'_{ij} \Sigma^{-1} \delta_{ij} = \frac{1}{\theta_1} (L'_1 \delta_{ij})^2 + \dots + \frac{1}{\theta_m} (L'_m \delta_{ij})^2. \quad (2.4.6)$$

Note that the eigen vectors provide a transformation of the original measurements into uncorrelated variables, in which case the Mahalanobis distance can be written as the sum of Mahalanobis distances due to different uncorrelated variables. We can choose any arbitrary set of vectors M_1, \dots, M_m such that $M'_i \Sigma M_j = 0$ for $i \neq j$ and $M'_i \Sigma M_i = 1$, to obtain a decomposition

$$\begin{aligned} D_{ij} &= \delta'_{ij} \Sigma^{-1} \delta_{ij} = (M'_1 \delta_{ij})^2 + \dots + (M'_m \delta_{ij})^2 \\ &= D_{ij}^{(1)} + \dots + D_{ij}^{(m)}. \end{aligned} \quad (2.4.7)$$

By combining some of the D_{ij} 's on the right hand side of (2.4.7), we obtain decompositions of D_{ij} with a smaller number of components.

If we choose

$$M_1 = (\sigma' \Sigma^{-1} \sigma)^{-\frac{1}{2}} \Sigma^{-1} \sigma \quad (2.4.8)$$

in (2.4.7), where σ is the vector of standard deviations of the individual characters (i.e., square roots of diagonal elements of Σ), then

$$(M'_1 \delta_{ij})^2 = D_{si}^2 \quad (2.4.9)$$

represents the component of Mahalanobis D^2 between π_i and π_j due to the size factor as defined by Rao (1962, 1971b).
Then

$$D_{ij}^2 = D^2 = D_{si}^2 + D_{sh}^2 \quad (2.4.10)$$

where D_{sh}^2 , the residual after subtracting the D^2 due to size, represents the distance due to shape factors between the two populations.

Penrose (1954) obtained a similar decomposition of Karl Pearson's CRL (coefficient of racial likeness) in terms of size and shape. The Penrose indices do not take into account the correlations that may exist between characters. For further details regarding the use of size and particular shape factors reference may be made to Rao (1962, 1971b).

2.5 Similarity Coefficients (SIMC's)

Instead of a difference measure between two individuals, it may be natural to consider a similarity function $s(X_1, X_2)$ and define S_i, S_j and S_{ij} by taking expectations analogous to H_i, H_j and H_{ij} . Then the DIVC of π_i may be defined by a suitable decreasing function of S_i , such as $1 - S_i$ or $-\log S_i$, specially when the range of S_i is $(0,1)$. The DISC obtained by choosing $H_i = 1 - S_i$ is

$$D_{ij} = \frac{1}{2}(S_i + S_j) - S_{ij} \quad (2.5.1)$$

and that by choosing $H_i = -\log S_i$ is

$$\begin{aligned} D_{ij} &= \frac{1}{2} (\log S_i + \log S_j) - \log S_{ij} \\ &= -\log \frac{S_{ij}}{\sqrt{S_i S_j}}. \end{aligned} \quad (2.5.2)$$

For instance, in the second example of Section 2.2, a natural definition of $s(X_1, X_2) = (\sum \delta_r)/m$, which lies in the range $(0,1)$. Then

$$S_i = J_{ii}, \quad S_j = J_{jj}, \quad S_{ij} = J_{ij} \quad (2.5.3)$$

where J_{ij} are as defined in (2.2.7), and using (2.5.1) and (2.5.2) we have the alternative forms

$$D_{ij} = \frac{1}{2} (J_{ii} + J_{jj}) - J_{ij}, \quad (2.5.4)$$

$$D_{ij} = -\log \frac{J_{ij}}{\sqrt{J_{ii} J_{jj}}}. \quad (2.5.5)$$

The expression (2.5.4) is the same as the "minimum genetic distance" (2.2.8) of Nei (1978), and (2.5.5) is what he calls the "standard genetic distance".

Again, in the example (2), we may define the similarity function as $(\delta_1 \dots \delta_m)^{1/m}$ instead of $(\delta_1 + \dots + \delta_m)/m$. The new function has the value unity when the gene alleles coincide at all the loci and zero otherwise. In such a case, when the characters are independent,

$$S_i = j_{ii}^{(1)} \dots j_{ii}^{(m)} = (J'_{ii})^m$$

$$S_{ij} = j_{ij}^{(1)} \dots j_{ij}^{(m)} = (J'_{ij})^m \quad (2.5.6)$$

where $j_{ij}^{(r)}$ are as defined in (2.2.5) and (2.2.6).

Taking logarithms of (2.5.6), the corresponding DISC is

$$D_{ij} = - \log \frac{J'_{ij}}{\sqrt{J'_{ii} J'_{jj}}} \quad (2.5.7)$$

which Nei calls the "maximum genetic distance".

2.6 A Functional Equation

Consider a multinomial distribution in k classes with probabilities $p = (p_1, \dots, p_k)$, and let $H(p)$ be a DIVC. The maximum DIV obtains when $p = (k^{-1}, \dots, k^{-1}) = e$, say (for evenness), so that we may have the condition:

$$C_1: \max_p H(p) = H(e). \quad (2.6.1)$$

Using $H(p)$ as a DIVC, we can construct a DISC between the multinomials defined by p and e by using (2.1.3),

$$D_{pe} = H\left(\frac{p+e}{2}\right) - \frac{1}{2} H(p) - \frac{1}{2} H(e). \quad (2.6.2)$$

The larger the value of $H(p)$, the closer p is to e , which suggests an alternative way of defining the DIS between the populations defined by p and e as a quantity proportional to

$$\max_p H(p) - H(p) = H(e) - H(p). \quad (2.6.3)$$

Equating (2.6.2) to a constant multiple of (2.6.3) we obtain the functional equation

$$H\left(\frac{p+e}{2}\right) - \frac{1}{2} [H(p)+H(e)] = c [H(e)-H(p)]$$

or

$$H\left(\frac{p+e}{2}\right) = \left(\frac{1}{2} + c\right) H(e) + \left(\frac{1}{2} - c\right) H(p). \quad (2.6.4)$$

where c is a constant. There may be many solutions to (2.6.4) subject to the condition C_1 . We shall impose some regularity conditions on $H(p)$ in order to restrict the solutions to a smaller class:

C_2 : $H(p)$ is symmetric in p_1, \dots, p_k

C_3 : $H(p)$ admits first and second order partial derivations with respect to p_1, \dots, p_{k-1} and the $(k-1) \times (k-1)$ matrix

$$H''(p) = \left(\frac{\partial}{\partial p_i} \frac{\partial}{\partial p_j} H(p) \right)$$

is continuous and not null at $p = e$.

Of course $H'(p) = 0$ at $p = e$ in view of the condition C_1 and the condition C_3 ensures that the diversity measure is locally sensitive when p deviates from e .

We shall show that under the conditions C_1 , C_2 and C_3 , the function $H(p)$ satisfying the equation (2.6.4) is of the form

$$H(p) = a (1 - \sum p_i^2) + b \quad (2.6.5)$$

where $a > 0$ and b are constants, i.e., $H(p)$ is essentially the Gini-Simpson index.

(i) Using the condition C_3 , we obtain on taking the first and second derivatives of both sides of (2.6.5) with respect to p_1, \dots, p_{k-1} ,

$$\frac{1}{2} H' \left(\frac{p+e}{2} \right) = \left(\frac{1}{2} - c \right) H' (p) \quad (2.6.6)$$

$$\frac{1}{4} H'' \left(\frac{p+e}{2} \right) = \left(\frac{1}{2} - c \right) H'' (p) \quad (2.6.7)$$

where H' is a $k-1$ vector and H'' is a $(k-1) \times (k-1)$ matrix. Putting $p=e$ in (2.6.7)

$$\frac{1}{4} H'' (e) = \left(\frac{1}{2} - c \right) H'' (e) \quad (2.6.8)$$

which implies that $c = 1/4$, using the condition $H''(e) \neq 0$

(ii) The equation (2.6.7) becomes

$$H'' \left(\frac{p+e}{2} \right) = H''(p). \quad (2.6.9)$$

Repeated use of (2.6.9) gives

$$H''(p) = H'' [2^{-n}(p-e)+e] \rightarrow H''(e) \quad (2.6.10)$$

The equation (2.6.10) implies that $H(p)$ is quadratic in p_1, \dots, p_{k-1} , which may be written, using the condition of symmetry,

$$\begin{aligned}
H(p) &= \lambda_1 \sum p_i^2 + \lambda_2 \sum \sum p_i p_j + \lambda_3 \sum p_i + \lambda_4 \\
&= \mu_1 \sum p_i^2 + \lambda_2 (1 - p_k)^2 + \lambda_3 (1 - p_k) + \lambda_4
\end{aligned} \tag{2.6.11}$$

where all the summations are taken from 1 to k-1. The condition C_2 demands symmetry with respect to p_1, \dots, p_k , in which case (2.6.11) assumes the form

$$H(p) = \mu_1 \sum_1^k p_i^2 + \mu_2. \tag{2.6.12}$$

Using the condition C_1 , we find that $\mu_1 < 0$ in which case $H(p)$ is of the form

$$a(1 - \sum p_i^2) + b \tag{2.6.13}$$

where $a > 0$, which is required to be proved.

3. ENTROPY AND INFORMATION

3.1 Measures of Entropy

A wide variety of DIVC's have been introduced through the concept of entropy and information. The general approach in these cases is basically different from that of Section 2.1, where a function $d(X_1, X_2)$ measuring the difference between individuals X_1 and X_2 is chosen first and probability distributions of X_1 and X_2 are used only to find the average of $d(X_1, X_2)$. In practice, $d(X_1, X_2)$ would be chosen to reflect some intrinsic dissimilarity between individuals relevant to a particular investigation. On the other hand, a measure of en-

entropy is directly conceived of as a function defined on the space of distribution functions, satisfying some postulates. Some of the postulates are that it is non-negative, attains the maximum for the uniform distribution and has the minimum when the distribution is degenerate. Thus a measure of entropy is an index of similarity of a distribution function with the uniform distribution, and hence a measure of DIV.

We shall consider the space of all multinomial distributions for simplicity of presentation of results, observing that the formulae for the continuous case can be obtained by replacing the summation by the integral sign. We represent the probabilities in the k cells of a general multinomial by p_1, \dots, p_k and for a particular population π_i by p_{i1}, \dots, p_{ik} . Mathai and Rathie (1975) consider three general forms for entropy:

$$H = (1 - \alpha)^{-1} \log (\sum p_r^{\alpha + \beta_r - 1} / \sum p_r^{\beta_r}) \quad (3.1.1)$$

$$H = [(\sum p_r^{\alpha + \beta_r - 1} / \sum p_r^{\beta_r}) - 1] \div (2^{1-\alpha} - 1) \quad (3.1.2)$$

$$H = - \sum p_r^{\beta_r} \log p_r / \sum p_r^{\beta_r} \quad (3.1.3)$$

where all the summations are taken from 1 to k . When $\beta_r = 1$ for all r we have the familiar expressions introduced by Renyi (1961), Havrda and Charvat (1967) and Shannon (1948).

All the functions (3.1.1) - (3.1.3) are non-negative, attain the maximum when p_i are equal (maximum diversity) and are zero when $p_i = 1, p_j = 0, j \neq i$ (minimum diversity). Mathai and Rathie (1975) discuss the various additional mathematical postulates which lead to these functions. Patil and Taille (1979) and Pielou (1975) provide interpretations of some of these functions in the context of ecological studies.

The functions (3.1.1) - (3.1.3) are all concave and the method of Section 2.1 can be used to construct a DISC between π_i and π_j . For instance, choosing (3.1.3) with $\beta_r = 1$ as a DIVC, and a mixture π_o of populations π_i and π_j with apriori probabilities λ_1 and λ_2 , we have

$$H_i = - \sum_{r=1}^k p_{ir} \log p_{ir}$$

$$H_o = - \sum_{r=1}^k (\lambda_1 p_{ir} + \lambda_2 p_{jr}) \log (\lambda_1 p_{ir} + \lambda_2 p_{jr}) \quad (3.1.4)$$

$$D_{ij} = H_o - \lambda_1 H_i - \lambda_2 H_j$$

$$= \lambda_1 \sum p_{ir} \log \frac{p_{ir}}{\lambda_1 p_{ir} + \lambda_2 p_{jr}} + \lambda_2 \sum p_{jr} \log \frac{p_{jr}}{\lambda_1 p_{ir} + \lambda_2 p_{jr}} \quad (3.1.5)$$

which is the information radius defined by Sibson and Jardine (1971) from other considerations.

Similarly, the DISC between π_i and π_j obtained by choosing (3.1.2) with $\beta_r = 1$ is

$$D_{ij} = [\Sigma(\lambda_1 p_{ir} + \lambda_2 p_{jr})^\alpha - \lambda_1 \Sigma p_{ir}^\alpha - \lambda_2 \Sigma p_{jr}^\alpha] \div (2^{1-\alpha} - 1) \quad (3.1.6)$$

which, when $\alpha = 2$, reduced to the Euclidean distance, apart from a constant multiplier,

$$2 \lambda_1 \lambda_2 \Sigma (p_{ir} - p_{jr})^2. \quad (3.1.7)$$

The DISC obtained by choosing (3.1.1) with $\beta_r = 1$ is

$$D_{ij} = \log \frac{\Sigma(\lambda_1 p_{ir} + \lambda_2 p_{jr})^\alpha}{(\Sigma p_{ir}^\alpha)^{\lambda_1} (\Sigma p_{jr}^\alpha)^{\lambda_2}}. \quad (3.1.8)$$

The formulae (3.1.5)-(3.1.8) involve explicitly the prior probabilities λ_1, λ_2 . In many practical applications, it is appropriate to choose $\lambda_1 = \lambda_2 = 1/2$ to define a DISC between two populations.

3.2 Apportionment of Diversity

By considering a mixture π_0 of populations π_1, \dots, π_m with prior probabilities $\lambda_1, \dots, \lambda_m$ we can obtain a decomposition of DIV in π_0 , based on any choice of the H functions (3.1.1)-(3.1.3),

$$\begin{aligned} H_0 &= \Sigma \lambda_r H_r + (H_0 - \Sigma \lambda_r H_r) \\ &= H(w) + D(b) \end{aligned} \quad (3.2.1)$$

as DIV within and DIS between populations. It may be noted that $D(b)$ cannot in general be obtained as a weighted combin-

ation of DISC's between all pairs of populations as in (2.3.1) for the choice of DIVC's derived by the method of Section 2.1. (It is, however, true when H is chosen as in (3.1.2) with $\beta_r = 1$ and $\alpha = 2$, in which case it also belongs to the class of DIVC's derived in Section 2.1). The ratio $G(b) = D(b)/H_0$ has been used by geneticists as an index of diversity between populations compared to within. However, as observed in Section 2.3, its value depends on the H function chosen. In their studies on diversity with respect to blood groups and biochemical markers, Lewontin (1972) used the H function (3.1.3) with $\beta_r = 1$, and Nei (1973) and Chakravarty (1974) used (3.1.2) with $\alpha = 2$ and $\beta_r = 1$. This raises the question as to what is the optimum choice of a DIVC in a given class $\{H\}$ to study the apportionment of DIV as between and within populations. A natural choice appears to be one which maximizes the ratio $G(b) = D(b)/H_0$ or minimizes the ratio $H(w)/H_0$. Such a choice will depend on populations under study and the prior probabilities.

To examine the extent to which the optimum choice depends on the population distributions, the following computations were made in the simple case of two binomial populations with equal prior probabilities. The class of H functions considered is a subclass of (3.1.2) and (3.1.3),

$$H^{(\alpha)} = (p_1^\alpha + p_2^\alpha - 1)(2^{1-\alpha} - 1)$$

where for $\alpha = 1$, the function is defined by the limiting value

$$H^{(1)} = - p_1 \log p_1 - p_2 \log p_2 .$$

Table 1 gives the values of $D(b)/H_0$ for different combinations of the proportions for the two binomials. For each combination, the first entry corresponds to the value of $G(b)$ for $\alpha = 1$, the second for $\alpha = 2$, the third for $\alpha = 2.5$, the fourth for the optimum α , and the fifth entry within brackets gives α_* , the optimum value of α . The blanks for certain combinations indicate that the values are the same as for the combination with the complimentary values of (p_1, q_1) , the binomial proportions of the two populations. It is seen that the optimum value α_* of depends on the values of p_1, q_1 , although it is stable for a wide range of values. If p_1 and q_1 are both small or both large α_* is small and tends to zero as p_1 and q_1 approach zero or unity. For values of p_1, q_1 near the boundary determined by the points $(.005, .7), (.01, .6), (.05, .5), (.1, .4), (.2, .3)$, α_* is close to unity which corresponds to the Shannon DIVC. For other ranges of (p_1, q_1) , α_* is nearly 2.5, although $\alpha = 2$, which corresponds to the Gini-Simpson index is a close competitor.

The values of the ratio $G(b)$ for the heptoglobin diversity in 25 Caucasian populations considered by Lewontin (1972) for different values of α are as follows:

| | | | |
|------------|-------|-------|-------|
| α : | 1.0 | 2.0 | 2.5 |
| $G(b)$: | .0209 | .0249 | .0251 |

The frequency of the heptoglobin allele in these cases varied between 21% and 45% except in one case it was 12%. The optimum α in such cases is about 2.5.

TABLE 1

Values of $G(b)$ Between Two Binomial Distributions Defined by p_1 and q_1

| p_1 | q_1 | .005 | .05 | .10 | .30 | .50 | .70 | .90 | .95 |
|-------|--------|--------|--------|--------|--------|--------|--------|--------|-----|
| .995 | .955* | .834 | .741 | .505 | .359 | .248 | .134 | .087 | |
| | .980 | .895 | .808 | .529 | .324 | .168 | .045 | .019 | |
| | .981 | .897 | .812 | .530 | .321 | .161 | .040 | .016 | |
| | .981 | .897 | .812 | .530 | .363 | .276 | .173 | .124 | |
| .95 | (2.43) | (2.44) | (2.45) | (2.45) | (0.80) | (0.60) | (0.54) | (0.54) | |
| | | .714 | .622 | .388 | .242 | .127 | .017 | | |
| | | .810 | .724 | .451 | .254 | .108 | .009 | | |
| | | .815 | .730 | .454 | .253 | .105 | .008 | | |
| .90 | | .815 | .730 | .454 | .255 | .127 | .019 | | |
| | | (2.44) | (2.45) | (2.45) | (1.69) | (0.97) | (0.73) | | |
| | | .531 | .305 | .167 | .065 | | | | |
| | | .640 | .375 | .191 | .063 | | | | |
| .70 | | .647 | .380 | .191 | .061 | | | | |
| | | .647 | .380 | .191 | .066 | | | | |
| | | (2.45) | (2.46) | (2.37) | (1.25) | | | | |
| | | .119 | .031 | | | | | | |
| | | .160 | .042 | | | | | | |
| | | .163 | .043 | | | | | | |
| | | .163 | .043 | | | | | | |
| | | (2.47) | (2.47) | | | | | | |

* The first four vertical entries correspond to $\alpha = 1, 2, 2.5$ and α_* respectively. The last entry within brackets is α_* , the optimal value.

Table 2 gives the values of H_0 and $G(b)$ for 9 blood group and 7 protein loci in the case of Makiritare Indians from 7 different villages. These were computed using the data kindly supplied by Chakravorthy (1974), assuming equal population sizes for the villages. It is seen from Table 2 that for the blood group loci, where p values are in the interval (30%, 70%), the optimum α is 2.5; and for the biochemical markers, where p values are in the interval (5%, 20%), the optimum α is 1, although the differences in G values are not large. The value of $\alpha = 2.5$ comes out better on the criterion suggested for the choice of a DIVC. However, the value of $\alpha = 2.0$ is a close competitor and has other desirable properties (see Burbea and Rao, 1980).

4. DISCRIMINATION INDEX

A general method of constructing DISC's is through the concept of discrimination between populations, i.e., the probability with which a given individual can be identified as a member of one of two populations to which he possibly belongs.

4.1 Overlap Distance (Rao, 1948, 1977; Wald, 1950)

Let X be a set of measurements which has the probability density $p_1(\cdot)$ in π_1 and $p_j(\cdot)$ in π_j . The best decision rule based on an observed value x of X , for discriminating between π_1 and π_j with prior probabilities in the ratio 1:1 is to assign x to

TABLE 2
Gene DIV of Makiritare Indians in Seven Villages
and Index of DIS Between Villages

| Locus | Ave p | $\alpha = 1$ H_O G(b) | $\alpha = 2$ H_O G(b) | $\alpha = 2.5$ H_O G(b) |
|------------------|----------|----------------------------|----------------------------|------------------------------|
| Serological | | | | |
| Diego | .196 | .7139 .1743 | .6303 .1711 | .6240 .1693 |
| Kidd | .336 | .9209 .0250 | .8924 .0320 | .8899 .0325 |
| Rh(C) | .418 | .9805 .0401 | .9731 .0542 | .9724 .0554 |
| P | .434 | .9874 .0172 | .9826 .0232 | .9821 .0237 |
| Lewis | .466 | .9967 .0791 | .9954 .1044 | .9952 .1191 |
| Ss | .470 | .9974 .0575 | .9964 .0770 | .9963 .0786 |
| Rh(E) | .563 | .9885 .0058 | .9841 .0079 | .9837 .0081 |
| MN | .714 | .8635 .0263 | .8168 .0291 | .8128 .0292 |
| Duffy | .736 | .8327 .0122 | .7772 .0142 | .7726 .0166 |
| Average | | .9202 .0415 | .8943 .0448 | .8921 .0486 |
| Biochemical | | | | |
| Ap | .0557 | .3101 .0647 | .2104 .0238 | .2054 .0213 |
| Hp | .424 | .9833 .0650 | .9769 .0866 | .9763 .0884 |
| Gc | .820 | .6801 .0431 | .5904 .0432 | .5837 .0427 |
| PGM ₁ | .848 | .6148 .0592 | .5156 .0504 | .5086 .0488 |
| Lp | .876 | .5407 .0084 | .4345 .0052 | .4275 .0047 |
| Alb | .9857 | .1081 .1293 | .0564 .1719 | .0547 .1444 |
| 6PGD | .991 | .0741 .2503 | .0357 .0678 | .0346 .0934 |
| Average | | .4730 .0561 | .4028 .0521 | .3987 .0522 |

$$\begin{aligned} &\text{population } \pi_i \text{ if } p_i(x) > p_j(x) \\ &\text{population } \pi_j \text{ if } p_i(x) < p_j(x) \end{aligned} \quad (4.1.1)$$

and to decide by tossing an unbiased coin when $p_i(x) = p_j(x)$.
The probability of correct classifications for the optimum
decision rule is

$$C_{ij} = \frac{1}{2} \int_{R_1} p_i(x) dx + \frac{1}{2} \int_{R_2} p_j(x) dx \quad (4.1.2)$$

where R_1 is the region $p_i(x) \geq p_j(x)$ and R_2 , the region $p_j(x) < p_i(x)$. The minimum value of (4.2) is $1/2$ which is attained when $p_i(\cdot) = p_j(\cdot)$, and the maximum is unity when the supports of $p_i(\cdot)$ and $p_j(\cdot)$ are disjoint. The more dissimilar the populations are, the greater would be the probability of correct classifications. Then we may define the DISC between π_i and π_j as

$$D_{ij} = C_{ij} - \frac{1}{2} \quad (4.1.3)$$

which is in the range $(0, \frac{1}{2})$. It is seen that

$$C_{ij} - \frac{1}{2} = \frac{1}{4} \int |p_i(x) - p_j(x)| dx \quad (4.1.4)$$

which is a multiple of Kolmogorov's variational distance or city block distance, which is a special case of the Minkowski distance

$$\left[\int |p_i(x) - p_j(x)|^t dx \right]^{1/t}, t \geq 1. \quad (4.1.5)$$

In the development of decision theory, Wald (1950) introduced the distance function between π_i and π_j

$$D_{ij} = \max_R \left| \int_R [p_i(x) - p_j(x)] dx \right| \quad (4.1.6)$$

where R represents any arbitrary region. The expression (4.1.6) is identifiable as

$$D_{ij} = 1 - \int \min[p_i(x), p_j(x)] dx \quad (4.1.7)$$

$$= \int_{R_1} [p_i(x) - p_j(x)] dx \quad (4.1.8)$$

where R_1 is the region $p_i(x) \geq p_j(x)$ as in (4.1.2). The expression (4.1.8) is the difference between the proportions of correct and wrong classifications by using the optimum decision rule (4.1.1). The expression (4.1.7) may be interpreted as the proportion of mismatched individuals in the two populations.

4.2 Quadratic Differential Metric (Rao, 1948)

Let us consider a family of probability densities $p(x, \theta)$, $\theta \in \Theta$, a k -vector parameter space. The Fisher information matrix at θ is $M = [m_{ij}(\theta)]$ where

$$m_{ij}(\theta) = \int \frac{1}{p} \frac{dp}{d\theta_i} \frac{dp}{d\theta_j} dx. \quad (4.2.1)$$

We endow the space Θ with the quadratic differential metric

$$\sum \sum m_{ij}(\theta) \delta \theta_i \delta \theta_j \quad (4.2.2)$$

and define the distance between two points θ_1 and θ_2 as the geodesic distance determined by (4.2.2). The expression (4.2.2) is a measure of difference between two probability distributions close to each other and the distance defined by it may be useful in evolutionary studies where gradual changes take place in a population in moving from state θ_1 to state θ_2 . In a recent paper Atkinson and Mitchell (1980) have derived the expressions for geodesic distances based on (4.2.2) for well known families of distributions.

4.3 Invariants of Jeffreys

Jeffreys (1948) defined what are called invariants between two distributions

$$I_m = \int | [p_i(x)]^{1/m} - [p_j(x)]^{1/m} |^m dx, \quad m > 0$$

$$I_0 = \int [p_i(x) - p_j(x)] \log \frac{p_i(x)}{p_j(x)} dx \quad (4.3.1)$$

where the second expression is the sum of Kullback-Leibler information numbers

$$I_{ij} = \int p_i(x) \log \frac{p_i(x)}{p_j(x)} dx, \quad I_{ji} = \int p_j(x) \log \frac{p_j(x)}{p_i(x)} dx. \quad (4.3.2)$$

When $m = 1$,

$$I_1 = \int | p_i(x) - p_j(x) | dx \quad (4.3.3)$$

which is Kolmogorov's variational distance (overlap distance of Rao, 1948). When $m = 2$

$$\begin{aligned} I_2 &= \int [\sqrt{p_i(x)} - \sqrt{p_j(x)}]^2 dx \\ &= 2 \left(1 - \int \sqrt{p_i(x) p_j(x)} dx \right) \end{aligned} \quad (4.3.4)$$

which is extensively used by Matusita (1957) in inference problems. The expression (4.3.4) is a function of the Hellinger distance

$$\cos^{-1} \int \sqrt{p_i(x) p_j(x)} dx. \quad (4.3.5)$$

Rao and Varadarajan (1963) have defined the Hellinger DISC to be

$$- \log_e \int \sqrt{p_i(x) p_j(x)} dx. \quad (4.3.6)$$

The measure (4.3.5) was proposed by Bhattacharya (1946) as a DISC between populations π_i and π_j and has been used in some genetic studies. The alternative expression (4.3.6) has an advantage over (4.3.5) in the sense that it is additive with respect to characteristics independently distributed in the populations.

It is seen that there are various approaches for measuring DIV and DIS and some of the controversies on the choice of these measures in practical investigations (see Li, 1978; Nei, 1978; Morton, 1973; and Smith, 1977) may be resolved through the concepts developed in the present paper. Some further work in this direction, which is in progress, will be reported elsewhere.

REFERENCES

- [1] Agresti, A. and Agresti, B. F. (1978). Statistical analysis of qualitative variation. Social Methodology (K.F. Schussler, Ed.) 204-237.
- [2] Atkinson, C. and Mitchell, A. F. S. (1980) Rao's distance measure. Sankhya (in press).
- [3] Bhargava, T. N. and Doyle, P. H. (1974). A geometric study of diversity. J. Theoretical Biology 43, 241-251.
- [4] Bhargava, T. N. and Uppuluri, V. R. R. (1975). On diversity in human ecology. Metron 34, 1-13.
- [5] Bhattacharya, A. (1946). A measure of divergence between two multinomial populations. Sankhya 7, 401.
- [6] Burbea, J. and Rao, C. R. (1980). On the convexity of Jensen difference arising from entropy of order α . Tech. Report. University of Pittsburgh.
- [7] Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. Amer. J. Hum. Genet., 19, 233-257.
- [8] Cavalli-Sforza, L. L. (1969). Human Diversity. Proc. XII Internat. Congr. Genetics, Tokyo 3, 405-16.
- [9] Chakravorthy, R. (1974). A note on Nei's measure of gene diversity in a substructural population. Humangenetik 21, 85-88.
- [10] Dennis, B., Patil, G. P., Rossi, O., Stehman, S. and Taille, C. (1979). A bibliography of literature on ecological diversity and related methodology. In Ecological Diversity in Theory and Practice 1 CPH, 319-354.
- [11] Gini, C. (1912). Variabilità e mutabilità. Studi Economico-aguridici della facotta di Giurisprudenza dell . Universite di Cagliari III, Parte II.
- [12] Havrada, J. and Charvat, F. (1967). Quantification method in classification processes: concept of structural α -entropy. Kybernetika 3, 30-35.
- [13] Jardine, N. and Gibson, R. (1971). Mathematical Taxonomy. Wiley, New York.

- [14] Jeffreys, H. (1948). Theory of Probability (2nd Edition). Clarendon Press, Oxford.
- [15] Karlin, S., Kenett, R. and Bonne-Tamir, B. (1979). Analysis of biochemical genetic data on Jewish populations: II. Results and interpretations of heterogeneity indices and distance measures with respect to standards. Am. J. Hum. Genet. 31, 341-365.
- [16] Lewontin, R. C. (1972). The apportionment of human diversity. Evolutionary Biology 6, 381-398.
- [17] Mahalanobis, P. C., Majumdar, D. N. and Rao, C. R. (1949). Anthropometric survey of the United Provinces, 1941: a statistical study. Sankhya 9, 90-324.
- [18] Majumdar, D. N. and Rao, C. R. (1958). Bengal anthropometric survey, 1945: a statistical study. Sankhya 19, 203-408.
- [19] Mathai A. and Rathie, P. N. (1974). Basic Concepts in Information Theory and Statistics. Wiley (Halsted Press).
- [20] Matusita, K. (1957). Decision rule based on the distance for the classification problem. Ann. Inst. Statist. Math. 8, 67-77.
- [21] Morton, N. E. and Lalovel, J. M. (1973). Topology of kinship in Micronesia. Amer. J. Hum. Genet. 25, 422-432.
- [22] Morton, N. E. (1973). Kinship and population structure. In Genetic Structure of Populations. pp. 66-71 (N.E. Morton, Ed.).
- [23] Nei, M. (1973). Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci., 70, 3321-3323.
- [24] Nei, M. (1978). The theory of genetic distance and evolution of human races. Jap. J. Human Genet., 23, 341-369.
- [25] Patil, G. P. and Taille, C. (1979). An overview of diversity. In Ecological Diversity in Theory and Practice. 1 CPH, 3-28.
- [26] Penrose, L. S. (1954). Distance, size and shape. Ann. Eugen. 18, 337-343.
- [27] Pielou, E. C. (1975). Ecological Diversity. John Wiley, New York.

- [28] Rao, C. R. (1945). Information and accuracy attainable in the estimation of parameters. Bull. Ca. Math. Soc., 37, 81-91.
- [29] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. J. Roy. Statist. Soc. B 10, 159-193.
- [30] Rao, C. R. (1954). On the rise and interpretation of distance functions in statistics. Bull. Int. Statist. Inst., 34, 90-
- [31] Rao, C. R. (1962). Use of discriminant and allied functions in multivariate analysis. Sankhya A 24, 149-154.
- [32] Rao, C. R. (1964). The use and interpretation of principal components in applied research. Sankhya A 26, 329-358.
- [33] Rao, C. R. (1965, 1973). Linear Statistical Inference and its Applications. John Wiley, New York
- [34] Rao, C. R. (1971a). Advanced Statistical Methods in Biometric Research. Haffner (first edition 1952). John Wiley, New York.
- [35] Rao, C. R. (1971b). Taxonomy in anthropology. In Mathematics in the Archaeological and Historical Sciences. 19-20. Edin. Univ. Press.
- [36] Rao, C. R. (1977). Cluster analysis applied to a study of race mixture in human populations. Proc. Michigan Univ. Symp. 175-197.
- [37] Renyi, A. (1961). On measures of entropy and information. Proc. Fourth Berkeley Symp. 1, 547-561.
- [38] Sanghvi, L. D. (1953). Comparison of genetic and morphological methods for a study of biological differences. Amer. J. Phy. Anthropol. 11, 385-404.
- [39] Sanghvi, L. D. and Balakrishnan, V. (1972). Comparison of different measures of genetic distance between human populations. In The Assessment of Population Affinities in Man. (J. S. Weiner and J. Huizinga, Eds.). pp. 25-36.
- [40] Sen, A. (1973). On Economic Inequality. Clarendon Press, Oxford.

- [41] Shannon, C. F. (1948). A mathematical theory of communications. Bell System Tech. J. 27, 379-423, 623-656.
- [42] Simpson, E. H. (1949). Measurement of diversity. Nature 163, 688.
- [43] Smith, C. A. B. (1977). A note on genetic distance. Ann. Hum. Genet. 40, 463-479. (London).
- [44] Sokhal, R. R. and Sneath, P. H. A. (1963). Principles of Numerical Taxonomy. W. H. Freeman.
- [45] Wald, A. (1950). Statistical Decision Functions. John Wiley, New York.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

(14) TR-80-10

| | | | |
|---|--|--|--|
| 18. REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM | |
| 1. REPORT NUMBER AFOSR/TR-80-0984 | | 2. GOVT ACCESSION NO. AD-A091033 | |
| 3. TITLE (and Subtitle) Diversity and Dissimilarity Coefficients: A Unified Approach. | | 4. TYPE OF REPORT & PERIOD COVERED Interim rept. | |
| 5. AUTHOR(s) C. Radhakrishna/Rao | | 6. CONTRACT OR GRANT NUMBER(s) F49620-79-C-0161 | |
| 7. PERFORMING ORGANIZATION NAME AND ADDRESS University of Pittsburgh Department of Mathematics and Statistics Pittsburgh, PA. 15260 | | 8. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2364A5 | |
| 9. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332 | | 10. REPORT DATE 11 July 1980 | |
| 11. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) (17) A5 (12) 36 | | 12. NUMBER OF PAGES 33 | |
| | | 13. SECURITY CLASS. (of this report) Unclassified | |
| | | 14. DECLASSIFICATION/DOWNGRADING SCHEDULE | |
| 15. DISTRIBUTION STATEMENT (of this Report) Approved for public release: distribution unlimited | | | |
| 16. DISTRIBUTION STATEMENT (of this abstract entered in Block 20, if different from Report) | | | |
| 17. SUPPLEMENTARY NOTES | | | |
| 18. KEY WORDS (Continue on reverse side if necessary and identify by block number) Entropy, Information, Diversity, Dissimilarity, Similarity, Mahalanobis D^2 , Size and Shape Factors, Discrimination, Geodesic Distance | | | |
| 19. ABSTRACT (Continue on reverse side if necessary and identify by block number) Three general methods for obtaining measures of diversity with- in a population and dissimilarity between populations are discussed. One is based on an intrinsic notion of dissimilarity between indiv- iduals and others make use of the concepts of entropy and discrimin- ation. A criterion is developed for choosing a measure of diversity in a given class of measures. The Gini-Simpson index of diversity is derived as the solution to a functional equation. | | | |